The Uses and Abuses of Psychometric Tests

William A. Hunt
Northwestern University

Follow this and additional works at: https://uknowledge.uky.edu/klj

Part of the Law and Psychology Commons

Right click to open a feedback form in a new tab to let us know how this document benefits you.

Recommended Citation
Available at: https://uknowledge.uky.edu/klj/vol35/iss1/2

This Article is brought to you for free and open access by the Law Journals at UKnowledge. It has been accepted for inclusion in Kentucky Law Journal by an authorized editor of UKnowledge. For more information, please contact UKnowledge@lsv.uky.edu.
THE USES AND ABUSES OF PSYCHOMETRIC TESTS

By William A. Hunt

HISTORICAL ANTECEDEENTS

The idea of estimating intelligence through the evaluation of a person's responses to a series of questions did not originate in a psychological laboratory. It has been a recognized judicial procedure for centuries. In 1534 Sir Anthony Fitzherbert writing in "The New Natura Brevium" said, "And he who shall be said to be a sot and idiot from his birth, is such a person who cannot account or number twenty pence, nor can tell who was his father and mother, nor how old he is, etc., so as it may appear that he hath no understanding of reason what shall be for his profit, nor what for his loss." Four hundred years later we find Alfred Binet, the father of intelligence testing, using the ability to count coins and the knowledge of one's age as items in his early intelligence test. The resemblance between
the questions asked by the psychologist on his intelligence tests and those asked by the court in establishing competency, as a witness has been commented on by Hutchins and Slesinger.\footnote{Hutchins, R. M., and Slesinger, D. \textit{Some Observations on the Law of Evidence—the Competency of Witnesses} (1928) 37 \textit{Yale Law Rev.} 1017.} Since both psychologist and judge are employing the same basic procedure, one may properly ask what advantage, if any, the psychologist’s laborious procedure has over the freer and easier one of the judge.

\textbf{Advantages of the Psychometric Approach}

We can classify the advantages of the psychological procedure under four headings—objectivity, reliability, standardization, and validity. These are the essential contributions of scientific methodology when applied to the evaluation of intelligence. In the first place, the psychologist asks relatively short, unambiguous questions the correct answers to which are immediately evident to anyone acquainted with the subject matter. This is what we call \textit{objectivity}. Thus, if we ask a subject how many pints there are in a quart, the correct answer is recognized at once without any further subjective evaluation of the response by the examiner. Objectivity is not always attained in all our test items. Thus when we ask a person, “Why are laws necessary?”, his response may vary and may be of different degrees of correctness. Here the examiner must decide how much credit to allow various types of answers. To help attain our goal of objectivity in this instance the examiner is provided with a series of typical responses and the standard credit allowed for each. These cover all of the common responses given, and if the examiner follows the instructions given him the personal element is largely removed from his evaluation.

This leads us to the second point. The same questions are always asked and asked in an identical fashion each time. This is \textit{standardization}. By always asking the same question in the same way, the examiner may compare the subject’s response with that of every other individual who is on record as having taken the test. The examiner does not have to rely upon his own experience with the test but may draw upon the recorded experience of other examiners. When the responses of large
groups of individuals are organized and published in order that they may be used for such purposes of comparison they are referred to as *norms*. Thus standardization, by rendering the response of any one individual comparable to that of all others on record, increases the objectivity mentioned above.

The third characteristic of the intelligence test is that it asks more questions than the court is apt to. Thus there are from 100-200 items on most of the current intelligence tests. The more questions we ask, the broader sampling we get of the individual's capacities, and the less weight is given to each item as a determinant in our final evaluation. Since extraneous, chance factors such as the misunderstanding of the question, inattention, or a temporary blocking in memory, may interfere to render some one response atypical of the subject's total capacity, the more items we introduce the less danger there is of distortion by such chance factors. This introduces the factor of *reliability* or the tendency of the test to give the same score when repeated upon another occasion. Within limits, reliability is a function of the number of items included in a test.

Finally, the psychologist always checks his test against other criteria before he accepts it as a measure of intelligence. The ability to count is not accepted as a typical manifestation of intelligence until it has been shown to increase with the increasing age of the child as we know intelligence increases; to be typical of "bright" children as opposed to "dull" children; and to be directly related to other accepted criteria of intellectual capacity. This agreement with other criteria is known as *validity*, the degree to which the test measures what it is supposed to measure. Such agreement with other criteria is never established by casual inspection or individual opinion but only after a careful empirical investigation governed by the established procedures of scientific research.

**PRACTICAL BASIS OF INTELLIGENCE TESTS**

Viewed in this light, it is evident that intelligence tests are not esoteric instruments arbitrarily devised within the cloistered atmosphere of some academic laboratory. They represent practical measures forged in the demands of everyday life and tempered by the application of rigid rules of scientific investigation. Their development is like that of any
test. First there arises a practical need of classifying people according to the amount of some important characteristic that they possess. It may be intelligence, it may be emotional stability, it may be the ability to perform clerical tasks; in any case its measurement seems useful to society. A group known to possess this characteristic is then selected and its behavior studied to ascertain whether or not there are any specific behaviors typical of this group which are not found in groups in which the characteristic in question is lacking. Thus, in regard to intelligence, we find that the average child of six or over can successfully count four pennies, while the average child below this age cannot. When such identifying behaviors that distinguish a group can be found and when repeated investigation shows that their appearance is related to the possession of the characteristic we are attempting to measure, we have a test. In the future when an individual behaves in this typical way, we infer that he possesses the trait in question. When we find a child who can count four pennies, we assume that in that respect at least his intelligence is equal to that of an average six-year-old or better. By adding to the number of such discriminative behaviors we improve the predictive value of our test.

There is nothing arbitrary here. Nothing is accepted that does not agree with experience. The mere fact that the psychologist has devised a strange vocabulary and many specialized statistical and experimental techniques for treating these problems should not blind us to the basic reasonableness of his approach and to the importance of his contribution. The practical, functional nature of intelligence tests stands out clearly when we survey the history of their development.

THE BINET-SIMON SCALES

By the beginning of this century the stage was set for the advent of intelligence testing. Experimental Psychology had devised techniques and procedures for the measurement of various psychological processes and Mathematics was beginning to contribute statistical formulae which could be utilized. Already sporadic attempts were being made to measure intelligence. Outstanding among the workers in this new field was a Frenchman, Alfred Binet, director of the Psychological
Laboratory at the Sorbonne in Paris. In 1904 the Minister of Public Instruction appointed Binet to a commission to study the problem of administering special classes in the public schools. One of the pressing problems before the commission was the devising of some efficient technique for the detection and classification of those children in need of special instruction. Here was a real opportunity for an intelligence test. Accordingly, Binet set to work and with his collaborator, Simon, published his first scale in 1905.

Binet's fundamental assumption was that intelligence is compounded of higher mental processes, such as attention, memory, etc., and that in measuring it we must approach the combined effects of several or all of them as exhibited in some single function. Thus attention and memory are both exhibited in the child's ability to repeat digits. In order to repeat three digits, he must both attend to them when they are spoken originally and remember them in order that he may reproduce them. For a test we must select many items in order to sample these faculties in various combinations and at various levels, as Galton says, by "the sinking of shafts at critical points." Binet also tried to use items that were relatively short, easily administered, and not directly dependent on school information. The items should increase in difficulty in order that the increases in intelligence as the child grows older may be differentiated. Among the items selected were the recognition of food, execution of simple commands, the naming of designated objects, repetition of digits, comparison of weights, drawing designs from memory, the completion of sentences, and the definition of abstract terms.

There were 30 items in all, arranged on the scale in order of their difficulty. The test was then administered to 200 normal children of varying ages. The performance of this group served as a standard of comparison. By consulting these rough norms, the examiner could tell how many items a child

---

2Binet, Alfred (1857-1911), French psychologist, director of the psychological laboratory at the Sorbonne. As an experimental and theoretical psychologist his early investigations of the bases of intelligence led to his construction of the first intelligence test.

3Galton, Sir Francis (1822-1911), an English scientist whose interest in individual differences and human faculties led him into the field of psychology.
of any given age might be expected to pass. Today we would consider Binet’s methods of validation primitive and his normative group as inadequate.

**The Concept of an Age Scale**

Binet continued working to improve his scale and published a revision in 1908. It was the first *age scale*, or test on which the items are grouped according to the chronological age for which the performance is typical. If from two-thirds to three-fourths of a particular age group passed an item, it was included among the items for that age level. If fewer than this passed it, it was put at a lower level. The subject’s performance is then reported in terms of his *mental age*. Mental age is the degree of intellectual development reached by the average child of any specific chronological age. If we say a child has a mental age of eight, we mean that his intelligence is that of the average eight-year-old child. On such age scales credit for passing the various items is given in terms of a certain number of months mental age for each item passed. The final mental age is attained by totalling the units of monthly credit. For the purposes of computation, some point is assumed as the subject’s basic mental age level. In Binet’s case it was that age level at which the child passes every item but one. Today it is customary to use the highest age level at which a child passes every item. To this basic value is added a proportional number of months credit for every item passed at a higher level. Thus, if a child passed every item at the eight year level, 4 out of 6 items at the nine year level, 2 out of 6 at the ten year level, we would call eight the basic level, add 4/6 of 12 months for his performance on the nine year items, and 2/6 of 12 months for the ten year items, ending up with a final mental age of 8 years plus 8 months plus 4 months, or 9 years. Such *age scales* may be contrasted with *point scales* where numerical rather than age credits are awarded. The numerical points are then totalled and mental age is obtained by comparison with a table of norms to see of what age group a certain point score is typical.

Binet revised his scale once more in 1911, just before his death. The average child of six is expected on this scale to be able to distinguish morning and afternoon, define words in terms of use, copy a diamond, count 13 coins, and make aesthetic
comparisons. At nine he should be able to make change, give
definitions superior to use, know the value of nine pieces of
money, repeat the months of the year, and comprehend simple
questions. At twelve he should be able to make a sentence using
three words given him, name 60 words in three minutes, define
abstract words, and arrange scrambled words into a sentence.
These are not esoteric, meaningless hurdles that the child is
arbitrarily put over, but are typical of the meaningful, intellec-
tual demands of his daily environment.

Terman's Revision of the Binet—
The Stanford-Binet Scale

American psychology, with its flare for the functional, im-
mmediately became interested in Binet's work. In 1910 Goddard⁴
published a revision for use in this country. With Binet's death
in 1911, the center of gravity of the testing movement shifted to
the United States and has remained here since. There have been
numerous subsequent revisions of the Binet scale with the names
of Kuhlmann,⁵ Yerkes,⁶ and Herring⁷ prominently connected
with the work, but the most widely used scale of all is the revision
by Professor Louis Terman⁸ of Stanford University. This first
appeared in finished form in 1916.⁹

Terman adapted the test items to the American scene, added
many more items, attempted to make them clear, and was more
careful in placing them in the proper age group. His scale
contained 90 items and was standardized upon the responses of

⁴Goddard, Henry Herbert (1866– ), Professor of Psychology at
Ohio State University, from 1906–18 director of research at the
Training School for Feeble-Minded Children, Vineland, N. J., an
authority on mental deficiency.

⁵Kuhlmann, F. A., Handbook of Mental Tests (Baltimore,
Warwick and York, 1922).

⁶Yerkes, R. M., Bridges, J. W., and Hardwick, R. S., A Point
Scale for Measuring Mental Ability (Baltimore, Warwick and
York, 1915).

⁷Herring, J. P., Herring Revision of the Binet–Simon Tests
(Yonkers-on-Hudson, World Book Company, 1931).

⁸Terman, Lewis Madison (1877– ), Professor of Psychology at
Stanford University, served in psychology division, Surgeon-
General's Office, during World War I. In addition to his work in
intelligence testing he is widely known for his studies of superior
children.

⁹Terman, L. M., The Measurement of Intelligence (Boston,
PsycHOMETRic Tests

some 2,000 subjects. It rapidly became the most widely used of all individual intelligence tests.

As time passed and experience with the Stanford-Binet Scale accumulated, certain difficulties became evident. A few specific items proved relatively unsatisfactory in use. The scale was weak at both age extremes. There was too much reliance on verbal factors. In some instances the objectivity of the items was lessened by ambiguity in their scoring. There also was a definite need for an alternate form to be used when it seemed desirable to retest a subject. Accordingly, in 1937 after a number of years’ work Terman published a revision of his scale in two equivalent forms, each form containing 129 items.¹⁰

This revision received a preliminary trial on 1,000 subjects who also had taken the original form. Its final standardization was based upon the performance of more than 3,000 subjects from 17 different communities in 11 states, carefully selected to cover a wide socio-economic range. This broad sampling was necessary in order that norms representative of the general population might be obtained. Its basic validation depends on the fact that the test scores show the expected increase as chronological age increases. In addition it also agrees well with other accepted criteria of intelligence such as school grades, teachers’ ratings, and the performance of known groups of superior children as well as institutionalized mental deficients.

The items remain of the familiar type we have seen before. The seven-year-old child is expected to give the number of his fingers, repeat a 16 syllable sentence, recognize simple pictorial absurdities, repeat 3 digits reversed, construct a sentence using three given words, and count and repeat 7, 5, and 8 taps made with a pen by the examiner. By the age of twelve he should be able to repeat designs, describe complex pictures, fill in missing words in a sentence, define abstract terms, recognize complex pictorial absurdities, and repeat 5 digits reversed. The average adult is expected to define abstract words such as generosity, justice, etc., getting 4 out of 5 correct; give opposite analogies; translate a simple code; explain proverbs; explain essential dif-

¹⁰TerMAN, L. M., and MERRILL, M. A., MeASURING INTEllIGENCE (Boston, Houghton Mifflin Company, 1937).
ferences, as between work and play; solve a paper cutting puzzle; and others.

Most psychologists agree that it is an improvement over the earlier scale. It still remains heavily verbal in nature, relying upon language in its administration and execution. Terman and Merrill defend this, saying: "Language, essentially, is the shorthand of the higher thought processes, and the level at which this shorthand functions is one of the most important determinants of the level of the processes themselves."\(^{11}\)

**Performance versus Verbal Tests**

Binet set as one of his original goals the development of items which were relatively independent of specific schooling. This motivation continues throughout the testing movement. Ideally we strive for some measure of inherent intellectual capacity. Actually we are limited always to measuring some behavior through which this assumed "innate" capacity expresses itself, and these behaviors are all affected by cultural experiences and schooling, some of them more than others. Fortunately, in a democratic society such as ours there is a wide diffusion of common cultural experiences and educational opportunities among all members of the society so that the inherent capacity of each child receives a roughly equal opportunity to develop and acquire cultural and educational knowledge. With opportunity equal, any differences in acquisition can be attributed to capacity to acquire, and we therefore try to limit our test items to those elements open to acquisition by all. Of course this is an ideal picture and there are many individuals and geographical and racial groups, whose educational opportunities and cultural backgrounds are relatively meager. Any test items unduly dependent upon schooling will handicap these people. Many psychologists feel that language is unduly influenced by schooling and that verbal items do not offer a true picture of the individual's capacities.

In an effort to avoid the weight attached to language by the use of verbal items, psychologists have devised *performance tests*, whose items involve the manipulation of concrete objects or pictures. There are many sorts of performance tests. In the well-known form-board type of test the subject is given flat

\(^{11}\)Terman, L. M., and Merrill, M. A., op. cit. supra, 10, at p. 5.
blocks of various geometrical shapes and required to fit them into correspondingly shaped holes in a board. Others involve the fitting together of the parts of a manikin or of a human profile. Sometimes simple picture puzzles are used. In the Kohs' Blocks Test the subject is asked to reproduce mosaic designs using a series of variously colored blocks. Often these tests involve printed materials with the subject requested to make pictorial absurdities or to designate similarities and differences. The common element in all such performance tests is their relative avoidance of the use of language.

THE INTELLIGENCE QUOTIENT

While the Stanford-Binet score is obtained directly in terms of mental age, it may then be translated into another measure, the Intelligence Quotient. Mental age gives us a measure of the subject's absolute level of capacity, but the Intelligence Quotient expresses his capacity in relation to that of other children of the same age. We know that a child with a mental age of eight is as intelligent as the average eight-year-old child. When we consider this performance in relation to the child's chronological age, however, further information accrues. If the child with a mental age (M. A.) of eight has the chronological age (C. A.) of eight, the child is normal, i.e. his mental level has kept an even pace with his chronological level. If his chronological age is ten, it is obvious that the child is retarded, i.e. his mental level has not kept pace with his chronological level. If his chronological age is only six, his intellectual development has obviously been accelerated. We express this relationship numerically by the Intelligence Quotient (I. Q.) according to the formula I. Q. = \( \frac{M.A.}{C.A.} \times 100 \). In the illustrations just given, where the mental age is divided by the chronological age and multiplied by 100, we would find for the normal child I. Q. = \( \frac{M.A.}{C.A.} \times 100 \) or \( \frac{8}{8} \times 100 = 100 \); for the retarded child I. Q. = \( \frac{M.A.}{C.A.} \times 100 = 80 \); for the superior child I. Q. = \( \frac{M.A.}{C.A.} \times 100 \) or \( \frac{8}{9} \times 100 = 133 \). I. Q.s of 100 then represent the average, I. Q.s below 100 represent retardation, and I. Q.s above 100 represent relative acceleration.

The following table gives a typical classification of I. Q.s
and the percentage of the population which might be expected to score at each level.\textsuperscript{12}

<table>
<thead>
<tr>
<th>Classification</th>
<th>I. Q.</th>
<th>% of Population Included</th>
</tr>
</thead>
<tbody>
<tr>
<td>Defective</td>
<td>65 and below</td>
<td>2.2</td>
</tr>
<tr>
<td>Borderline</td>
<td>66 - 79</td>
<td>6.7</td>
</tr>
<tr>
<td>Dull Normal</td>
<td>80 - 90</td>
<td>16.1</td>
</tr>
<tr>
<td>Average</td>
<td>91 - 110</td>
<td>50.0</td>
</tr>
<tr>
<td>Bright Normal</td>
<td>111 - 119</td>
<td>16.1</td>
</tr>
<tr>
<td>Superior</td>
<td>120 - 127</td>
<td>6.7</td>
</tr>
<tr>
<td>Very Superior</td>
<td>128 and over</td>
<td>2.2</td>
</tr>
</tbody>
</table>

It should be noted that the Intelligence Quotient is merely a rough index number for expressing relative intelligence. Differences of one or two points are by no means significant and should not be evaluated as such.

\textbf{THE INTELLIGENCE QUOTIENT IN ADULTS}

The intelligence quotient is spurious when used with adults. Intelligence increases with age only to a certain point. This is usually assumed to be somewhere between 14 to 20 years of age, with the major growth probably over by 16. Since it is unfair to expect the adult intelligence to increase year by year to balance the constantly increasing chronological age in the $\frac{M.A.}{C.A.}$ formula, we use a single fixed mental age (usually 15, although 14 or 16 may be used) as the denominator in our fraction when dealing with adults. In the case of the child, the I. Q. expresses his development in relation to all other children of the same chronological age. Since a single chronological age (15) is used in computing the I. Q. for adults, the I. Q. expresses the subject's intelligence relative to the entire adult population, rather than to the members of any single age group. All such numerical expressions are highly arbitrary, and it should be remembered that their authority stems ultimately from the actual performance of the subject in dealing with the specific test items.

\textbf{THE WECHSLER-BELLEVUE SCALE}

Even in its latest revision the Stanford-Binet Scale is best adapted to testing children. To fit the needs of institutions

where adults must be tested, Dr. David Wechsler of the Bellevue Psychiatric Hospital has developed the Wechsler-Bellevue Scale. While Wechsler’s philosophy of testing differs little from the traditional approach, he has constructed his test from items designed for adults, and has validated it upon an adult group. He has included more performance items, one-half of the test consisting of such material. An example of this is the Picture Arrangement test which consists of several series of individual pictures. The pictures in each series are presented to the subject in a haphazard order. When they are arranged in proper sequence, however, they tell a story. This problem involves no overt language manipulation by the subject. Moreover, in selecting the story-telling sequences, Wechsler has drawn upon the amusing and sophisticated adventures of Soglow’s Little King, who will be remembered from the pages of the New Yorker. Such material is calculated to arouse the interest and cooperation of adult subjects.

The Wechsler-Bellevue Scale differs from the Stanford-Binet in two other important ways. In the first place, the items are not segregated according to age level, but according to the type of mental function they are assumed to sample, with all the items sampling information collected in a group, thus forming a sub-test for range of general information. In the same fashion comprehension, memory, ability to abstract essential similarities, arithmetic, etc., are all represented by sub-tests the items of which all measure a specific intellectual function. This makes possible an easy comparison of the different functions sampled by each sub-test, a comparison which is achieved only with difficulty when the items are organized by age level rather than by type. Age level organization does not lead to efficient analysis of function, as we can see by going back and looking at the varied nature of the items included at any age level of the Binet. This confusion has resulted in the Stanford-Binet sometimes being referred to as a “hotchpotch.” In the second place, the Wechsler-Bellevue is a point scale rather than an age scale. Credits are assigned each correct response, not in terms of so many months of mental age but in terms of a certain number of numerical points. These points are then added to give a raw

---

13 Wechsler, D., op. cit. supra, n. 12.
score which can be turned into either mental age or intelligence quotient by consulting a table of norms (equivalent scores).

The Wechsler-Bellevue was finally standardized on a group of 1600 subjects. The original sampling involved the testing of approximately 3500 people, the majority of whom were adults. Not only do scores on the test show the same distribution over the various chronological age levels that intelligence does, but they show good agreement with other known criteria for intelligence.

The Coefficient of Correlation

The psychologist expresses the amount of such agreement by an index figure which he calls the coefficient of correlation. This value which may be obtained by various mathematical formulae expresses co-relationship, or the tendency of two factors to appear together. It ranges from +1.0 through 0 to —1.0. A correlation coefficient of +1.0 indicates a perfect positive correlation and means that the two factors being studied always appear together so that the presence of one implies the presence of the other. A coefficient of 0 means that there is no relationship between the two and that the presence of one leads to no prediction concerning the presence of the other. A value of —1.0 indicates a perfect negative relationship with the presence of one factor always implying the absence of the other. Wechsler has compared the performance of subjects on his test with their performance on the Stanford-Binet and finds a coefficient of correlation of +.82 indicating that the tests are in good agreement and therefore presumably measuring the same thing. That this "thing" is intelligence is further indicated by the finding that there is a correlation of +.79 between the performance on the Wechsler-Bellevue Scale and psychiatrists' recommendations for institutional commitment for mental deficiency. In this study involving the test's ability to predict the psychiatric judgment the Wechsler Scale outstripped the efficiency of the Stanford-Binet. Finally, the Wechsler Scale agreed +.43 and +.52 with teachers' ratings in school. These last correlations are of the size that wide experience has led us to expect between intelligence and school performance where other factors such as motivation enter to determine a student's grades. It would seem proper to conclude that the Wechsler Bellevue is a serviceable test of intelligence.
DESCRIPTION OF THE WECHSLER-BELLEVUE SCALE

The Wechsler-Bellevue Scale in its final form consists of 5 verbal tests and 5 performance tests. There is also an alternate vocabulary test which may be used if the examiner desires. When we find a scale divided into sub-tests in this fashion, we sometimes speak of the entire scale or group of sub-tests as a battery. In view of the importance of this scale and its wide use we will describe it at some length. The 5 verbal tests are designed to measure information, comprehension, arithmetic, the ability to recognize essential similarities between two things, and the ability to repeat a given series of digits, both forward and backward. Information and Arithmetic are completely objective and are scored on a simple basis of right or wrong.

Here are some of the typical questions from the sub-test for information:

- Who is the President of the U. S.?
- Who was the President before him?
- How many pints make a quart?
- What is a thermometer?
- How many weeks are there in a year?
- What is the capital of Italy?
- When is Washington's birthday?
- Where is Brazil?
- Who wrote Hamlet?
- What is the capital of Japan?
- What does the heart do?
- What is the population of the U. S.?
- What is the Koran?
- What is an Habeas Corpus?
- What is the Apocrypha?

Here are some of the arithmetic items:

- If a man buys six cents worth of stamps and gives the clerk ten cents, how much change should he get back?
- If a man buys eight cents worth of stamps and gives the clerk twenty-five cents, how much change should be get back?
- How many hours will it take a man to walk twenty-four miles at the rate of three miles an hour?
- If a man buys seven two cent stamps and gives the clerk a half dollar, how much change should he get back?
A man bought a second hand car for two-thirds of what it cost new. He paid $400 for it. How much did it cost new?

Eight men can finish a job in six days. How many men will be needed to finish it in a half day?

**Comprehension** allows for some range of quality in the answers and each correctly answered question is allotted either one or two points, depending upon the degree of excellence and generalization of the subject's answer. Thus the first question, "What is the thing to do if you find an envelope in the street, that is sealed, and addressed and has a new stamp?" could be correctly answered by responding "Take it to the police." However, the answer "Drop it in a mail box" is obviously superior and is allotted two points where the first answer would only receive one. While Wechsler has included sample responses to guide the examiner in scoring, some subjectivity is introduced when the examiner is forced to evaluate a novel response not found in the scoring tables.

Some of the comprehension items follow:

- What should you do if while sitting in the movies (theatre) you were the first person to discover a fire (or see smoke and fire)?
- Why are shoes made of leather?
- Why does land in the city cost more than land in the country?
- If you were lost in a forest (woods) in the daytime, how would you go about finding your way out?
- Why are people who are born deaf usually unable to talk?

In **Similarities** the subject is presented with the names of two things and asked to tell in what way they are alike. There are twenty pairs of names ranging in difficulty from an orange and a banana to a fly and a tree. Here again some subjectivity enters into the scoring since the degree and quality of the subjects' generalizations may vary. The answer that an orange and a banana are alike because both have peels receives only 1 point, while two points are given for the answer that they are both fruits. The answer that a fly and a tree are alike because they both grow receives 1 point, while the answer that both of them are living things receives 2 points.

In **Digit Repetition**, the examiner speaks a series of numbers and the subject repeats the series when the examiner has finished. The series start with 3 digits and increase gradually to 9. When the individuals' limit for forward series is reached,
the examiner begins again with another group this time request-
ing the subject to repeat the digits backward.

The performance portion of the Wechsler battery consists of
five sub-tests, Picture Completion, Picture Arrangement, Object
Assembly, Block Design, and Digit Symbol. Picture Completion
consists of a series of fifteen pictures of objects each of which
has some part missing, and the subject is supposed to identify
the missing part. The pictures include such things as the face
of a girl with the nose absent, a man’s face with only half a mus-
tache present, a door without a knob, an electric light bulb with
the thread missing from the base, and the figure of a man stand-
ing in bright sunlight without casting a shadow.

Picture Arrangement contains six separate series of pictures,
each series consisting of from 3 to 6 individual pictures which
tell a story if arranged in proper sequence. The incidents are
pitched at a humorous, sophisticated level to catch the interest
and insure the cooperation of adults.

Object Assembly consists of 3 items, a jig-saw mannikin
whose head, arms, legs, and body are separate and must be as-
sembled correctly; a human profile with the features cut out
in various shaped pieces to be reassembled; and the profile of a
hand, also cut up and to be put together again. These resemble
a simple puzzle situation.

Block Design is an adaptation of the Kohs’ Blocks Test.
The subject is given a number of vari-colored blocks which he
must assemble to duplicate a series of mosaic patterns given him.
The designs range from simple patterns involving only 4 blocks
up to complicated ones demanding 16 blocks.

The Digit Symbol Test resembles a simple code substitution
test. The subject is presented with a key consisting of a series
of numbers each of which is paired with its own particular sym-
bol. He then has to go through another lengthy series of num-
bers, writing beneath each number the appropriate symbol as
designated in the key. In this test, as in all the performance
tests, the subject works against a time limit, and the final score is
based not only upon the correctness of the solution but also upon
the length of time taken to achieve it.

The Vocabulary Test simply checks the subject’s ability to
define a list of 50 words, ranging from such simple ones as
“apple” and “donkey” through “stanza” and “microscope”
to the difficult "flout" and "traduce." It is used as an alternate test and is not regularly given as a part of the scale.

Scatter

Since all those related items which measure the same aspect or dimension of intelligence are grouped together in separate sub-tests in the Wechsler-Bellevue Scale, it is possible to establish independent norms for each sub-test and thus separately evaluate the different types of intellectual performance as well as intelligence as a whole as represented by the total score on the complete battery. In general, most individuals do not show much variation in performance level between the different sub-tests, and we may expect the ordinary subject's vocabulary, memory, abstracting ability, etc., all to be roughly comparable. Sometimes, however, we may find wide discrepancies, or scatter, between the scores on the various parts of the test. Since such extreme scatter is unusual, we assume the presence of some unusual causal factors in the individual's development. Occasionally an environmental factor such as lack of schooling may be responsible, as when we find that an educationally handicapped individual may perform much better on a performance test such as Block Design than he will on verbal tests such as Information and Arithmetic. More often in the people usually handled in psychological clinics such scatter may be indicative of the presence of mental disorder. This scatter is attributable to the fact that in many of the psychoses (major mental illnesses) and organic brain conditions there is a progressive uneven deterioration in the level of the individual's intellectual ability which does not affect all the components of intelligence equally. In these disorders we speak of dementia, meaning that the individual has slipped back below a previously attained level of intelligence. He shows a loss of intelligence, a deterioration, or we might say a psychological deficit. Dementia is thus opposed to amentia or primary mental deficiency as seen in the typical feebleminded individual who has never attained a normal level of intellectual functioning and hence shows a lack rather than a loss of intelligence.

The deterioration or deficit typical of mental disorder is not apt to affect all the components of intelligence equally and the deficit will be spotty and irregularly distributed among the
various sub-tests of our intelligence test battery. We frequently find, as we do in senility, that the older, well-established memories and habits remain relatively less affected than more recent ones. This may show itself in a vocabulary test score which is higher than that for the repetition of digits, since language represents an old habit ingrained by long use whereas the repetition of digits is based upon recent memory. In other disorders, the ability to handle abstractions, or to think in the abstract, as demanded by the Similarities test may suffer more damage than the ability to think concretely as exhibited in the more or less rote memory performance demanded by the Information test. Wide discrepancies of this type are suggestive of mental disorder either of an organic type (i.e. with structural damage to the nervous system, especially the brain) or of a functional type (i.e. mental illness without underlying structural damage.) On the assumption that different mental disorders may show different differential effects upon the various components of the total intellectual process, we then hope to use such differential effects as a diagnostic index to the mental disorder present. When used for such diagnostic analysis, the intelligence test is not made use of as an index of intellectual level but rather as a diagnostic index of mental disorder. Such diagnostic use of intelligence tests forms an important aspect of current clinical psychological practice. Here again, the basic assumption is that our intelligence test will reveal the intellect in action, and that unusual peculiarities in such functioning are diagnostic of some unusual conditions in the individual, in this case mental disorder. It should be pointed out, however, that at the present stage of our knowledge the clinical signs revealed in such a fashion are indicative rather than diagnostic and must be interpreted with great care, since other factors such as educational and cultural handicap, embarrassment, anxiety, and even a lack of rapport between the tester and the subject may produce such scatter.

THE DETERIORATION OF INTELLIGENCE WITH AGE

In discussing the use of the intelligence quotient, we mentioned that intelligence is assumed to mature sometime during late adolescence, which makes it necessary to use a constant chronological age in computing the intelligence quotient of all
adults. Thus, while an intelligence quotient places a juvenile in relation to other members of his particular age group, the intelligence quotient of an adult establishes his performance in relation to all other adults without reference to possible differential factors of age. The assumption that once intelligence matures it remains constant is not correct, and we know that somewhere in the middle twenties the intellect having matured then begins slowly to deteriorate. Such deterioration accelerates somewhere around fifty to sixty years of age so that a subject sixty years of age with an original intelligence quotient of 100 will not perform as capably as a younger man of the same original intellectual level. There have been many arguments in the literature as to whether this deterioration represents a true deterioration in ability or power or merely reflects a deterioration in the speed with which intelligence acts, but everyone is agreed that some such type of deficit appears. Fortunately, the older person usually is able to compensate for this deficit in functional potentiality by his years of acquired experience. One of the peculiar advantages of the Wechsler-Bellevue test stems from the fact that separate age norms have been established for the various adult age groups, so that it is possible in scoring the Wechsler-Bellevue to take this age deficit into account. An intelligence quotient computed from this scale thus compares the adult with other adults in a similar age range, and does not penalize the older individual by comparing his performance with that of younger men.

**Group Tests**

So far we have been talking about *individual* tests of intelligence where the test is administered personally by the psychologist to a single subject. Such individual testing enables the psychologist to give close attention to the subject's responses and to supplement the test score by personal observation of the subject's total behavior in the testing situation. Unfortunately the giving and scoring of such individual tests, however satisfactory they may be, is exceedingly laborious and time-consuming and therefore not suitable when large numbers of people must be tested in a relatively short time. For this latter purpose *group* tests have been devised.
A careful examination of the test items that we have quoted above will reveal that most of them could be presented as efficiently to the subject by printing them on paper and having him read them and write down his answer. Information questions, similarities, arithmetic, etc., all lend themselves to reproduction in a printed booklet which can also provide space for writing in the answers. Performance material is not as easily translated into a printed test but can be done in picture fashion. Thus the Revised Minnesota Paper Form Board\(^4\) is devised to measure the subject's ability to handle spatial relationships. Here the subject is presented with the outline of a complete geometric figure. At the same time he is given five alternative sets of geometric parts and asked to choose that set which, if assembled together, would give the total design presented in the outline.

Not only can intelligence test items be printed in booklet form but further objectivity can be obtained by stylizing the form of the individual's responses. Instead of allowing him to write out an answer in long hand in his own way, he may be required to answer with either a single letter, number, or check mark made upon the test paper. This makes it possible to place a scoring stencil over the subject's test paper and quickly check the number of correct answers. It also is possible to present the questions in a test booklet and to provide separate special answer sheets, the answers on which can be scored mechanically. This recent development makes possible the use of various types of standard business machines for handling large numbers of papers with extreme rapidity. Speed is not the only factor of advantage in this process, as the automatic scoring of the machine will eliminate the occasional human errors which creep in if the tests are scored by hand. In this manner, complete objectivity and standardization in the evaluation of test performance is carried close to ultimate perfection.

**The Army Alpha**

It was not long after the introduction of individual testing that the possibility of adapting the technique to group testing began to be realized. By the outbreak of hostilities in World

---

\(^4\)Revised Minnesota Paper Form Board, distributed by the Psychological Corporation, New York.
War I, A. S. Otis\textsuperscript{15} was already preparing a group test for publication. The tremendous need for selective placement in the military services gave a great impetus to the development of group testing. The military authorities badly needed some means of segregating the mentally incompetent, and classifying other men according to their mental capacity in order that they might be placed in positions commensurate with their intellectual ability. In April, 1917, a committee of the American Psychological Association under Professor Robert M. Yerkes\textsuperscript{16} was appointed to attack the general problem of preparing a group test or tests suitable for military use. Such a test not only had to meet the requirement of evaluating large numbers of individuals in rapid fashion; but had to be suitably graded in difficulty to allow for the large range of ability being tested; had to be simple, rapid, and objective so that it could be handled with a minimum of error; and had to be relatively independent of specific school information; as well as involving a minimum of writing. It was also necessary to provide a number of alternate forms to avoid coaching and malingering. The result was the famous Army Alpha Group Test of Intelligence, the prototype of all subsequent group tests. The Army Alpha, printed in booklet form, consisted of eight tests which measured the ability of the individual to follow simple directions, to handle arithmetical problems, to exercise practical judgment, to demonstrate language ability in the handling of synonyms and antonyms, to rearrange disarranged sentences, to follow through incomplete number series, to handle general analogies, and to display his range of general information. The test was tried out on large numbers of Army recruits and validated against school grades, teachers' estimates of ability, other intelligence tests, officers' rating, rank attained, ability shown during basic military training, previous civilian attainments, etc. All of these are commonly accepted criteria of intellectual attainment. The final test was given with great success to over a million recruits. While there is much about the Army Alpha that we would criticize

\textsuperscript{15} Otis, A. S. (1886-- ), Consultant, Division of Research and Test Service, World Book Company; psychological consultant to the military services in both World Wars.

\textsuperscript{16} Yerkes, R. M. (1876-- ), Professor of Psychology at Yale University, and formerly director of their primate laboratory, chief of the psychology division, Surgeon General's Office, during the first World War.
today, it filled a genuine need in World War I and was an outstanding example of the practical value of intelligence tests when applied to the problems of selective placement.\textsuperscript{17}

**The Army Beta**

Since the problem of illiteracy loomed large in the Army, (it was estimated that thirty per cent of the recruits were not able to handle the Army Alpha because of reading or language handicaps) it was necessary to devise a non-language pictorial group test for use in such cases. The result was the Army Beta, again the classical prototype for non-language paper and pencil tests. It consisted of such sub-tests as a maze test in which the individual was presented with a set of printed mazes in which he traced his route to the exit by pencil, a digit symbol test which is essentially a test of his ability to learn a simple code, a number checking test in which two columns of numbers are compared for similarity or difference, a set of incomplete pictures in which the subject checked the missing part, and other such essentially pictorial tests, all printed in the booklet form typical of the Army Alpha.\textsuperscript{18}

**The Later Development of Group Tests**

The outstanding success of the Army intelligence tests resulted in a tremendous boom in psychological testing following World War I. Both the original over-optimism of psychologists and the resultant public resistance to testing were gradually overcome and group intelligence testing came of age during the late twenties and early thirties. A classical example of such group tests is the Otis Self-Administering Test of Mental Ability\textsuperscript{19} which served as a model for the General Classification Test which was in use by the Navy as a standard selection measure for recruits by the outbreak of World War II. Such group tests of intelligence were so thoroughly developed by this time that despite their wide use in this last war only minor changes in their form resulted. A wide variety of well standardized tests is now available for anyone who cares to use them.

\textsuperscript{17} YOAKUM, C. S., and YERKES, R. M., ARMY MENTAL TESTS (New York, Henry Holt and Company, 1920).

\textsuperscript{18} YERKES, R. M. (Editor), Psychological Examining in the United States Army, NATIONAL ACADEMY OF SCIENCES, vol. XV, Washington, 1921.

\textsuperscript{19} Otis Self-Administering Tests of Mental Ability, distributed by World Book Company, Yonkers-on-Hudson.
Group tests are essentially point scales, that is, credit for each item answered correctly is not assigned in terms of some developmental unit such as months of mental age but in simple mathematical units, say 1 point of credit for each item passed. The total number of points is then added and this raw score is compared with the scores of a representative sampling of the population for which other measures such as mental ages and intelligence quotients already have been determined. The point scale score may then be translated into mental age or intelligence quotient. Many psychologists, however, protest that these last two measures are only valid for interpreting the individual's performance on an age scale test, such as the Stanford-Binet, and are not valid for use with the results of group tests.

Another measure is the centile score. A centile (percentile) may be defined as that point on the total distribution of scores below which fall any given percentage of the population tested. To assign an individual a centile score of 93 would mean that 93% of the population in question did more poorly on the test than he did. A centile score of 3 would mean that only 3% of the normative group performed more poorly. The centile score thus gives us another means of placing the individual in relation to the group. There are other methods of doing this, most of them derivations of the standard score, which is essentially a means of evaluating the individual's performance in terms of the amount by which he deviates from the average performance of the group. This amount is expressed in some unit of dispersion on the distribution curve established for the group. Its mathematical intricacies have no place here.

In validating any intelligence test and in setting up standard norms for the interpretation of scores on the test it is of course impossible to test the entire population at large. Instead we must select a limited sampling from the whole population and use this in our work. If we select a sample which is truly representative, then our test and the norms of performance that we

---

20 See p. 8.
establish for it are assumed to be adequate for the total population from which the sampling was drawn. We have seen how Terman standardized the revised Stanford-Binet upon 3,000 subjects from 17 different communities in 11 States assuming that this sampling was broad enough to be representative of the general population in the United States. It is well to remember, however, that this is merely an assumption. As we test more and more people and add the results to our distribution of scores, the norms we derive from this distribution become increasingly typical of the entire population. The goal of perfect representation is still to be reached, and one of the ways of improving our tests is to increase the size and representative quality of the samplings upon which they are based. Because of this some psychologists prefer to view test norms, not as general norms representative of the population at large, but as special norms representative only of the group from which they were derived. There are also many situations in which it is desirable to compare the individual with some special group. It may be important to know not how a child's intelligence compares with all other children but how it compares with that of the other children in his class. Success in college may depend not on how a student's intelligence compares with that of all other college students but on how well he approximates the average performance of the students at his particular institution. Consequently there is an increasing stress on local norms for specific groups. The stress here is not upon the position of the individual within the general population but upon his relative position in the particular group within which he is functioning at the moment.

THE USES OF INTELLIGENCE TESTS

Intelligence tests are useful wherever a reliable measure of an individual's intelligence is desired. For the subjective, impressionistic judgment which was previously relied upon in evaluating this important aspect of human behavior they offer a substitute which is objective, scientific, and relatively fool-proof. While it is not possible to predict success in any given endeavor upon the sole basis of the intelligence of the individual concerned, we do know that there are levels of intelligence, depending upon

---

*See p. 10.*
the educational or vocational task in question, below which an individual cannot fall and still succeed. There is a minimum level of intelligence which is necessary if the person is to have a chance at success. While such selection is negative, that is, it weeds out certain individuals who are bound to fail, but does not predict that the individuals above this level are bound to succeed, it has an important placement function in education, industry, and in the military services.

Another use of intelligence tests may be called clinical. We know that intelligence is an important factor in the ability of the human individual to adjust successfully to his environment. Many cases of maladjustment are attributable, in whole or in part, to the fact that the maladjusted individual is not sufficiently intelligent to cope successfully with the demands made upon him by some life situation. Here an adequate measure of the individual's intelligence may give us insight into the genesis of his emotional problems. Both the placement and clinical uses of tests may be combined in some instances. A factory worker may show a proneness to accidents, or may develop an emotional upset reflected in a high rate of absenteeism, if he is called upon to perform at a level which is above the potentialities of his intellect. An intelligence test will thus both give us clinical insight into the genesis of his difficulties and guide us in placing him in a position more commensurate with his abilities. Intelligence tests have had their widest application in the educational field, but have also been applied in industry, the military services, psychiatric practice, and the law.

**Intelligence Tests in Placement**

We have seen that Binet's work on his original intelligence test derived its practical motivation from his desire to aid the school authorities in Paris in selecting and training those pupils who were in need of special classes. Such selection of pupils for special treatment remains today one of the primary contributions that intelligence tests make in our educational system. By this means, those students of low intelligence who are not able to follow the pace of their more normal classmates can be identified and isolated for enrollment in special classes. There is another side to such testing—many pupils who fall behind in

\[^{23}\text{See p. 6.}\]**
their class work and are presumed to be lacking in intellect demonstrate normal ability on such tests. This makes possible a more accurate analysis of the child’s problem and often reveals the real source of scholastic difficulty as some behavior problem, physical handicap such as deafness or poor visions, or deficiency in some special ability such as reading. Special classes for the mentally retarded have in turn been followed by special classes for the superior child, and intelligence tests make possible the identification of those pupils whose superior intellect enables them to benefit by a faster rate of progress in the classroom.

At higher educational levels, such as college and professional schools, preliminary testing is sometimes resorted to as a means of eliminating candidates for higher learning who cannot be expected to meet the high level of scholastic accomplishment which is demanded. We know that while intelligence and scholastic accomplishment are inter-related intelligence is not the only factor operating in educational achievement, particularly at the college level. The tests used in collegiate selection therefore usually are not spoken of as intelligence tests, but as scholastic aptitude tests, in recognition of the fact that they may be measuring other factors besides intelligence. Previous educational and cultural experience also determine scholastic performance. Such tests are validated directly against the ability of the individual as shown in his subsequent class performance, rather than against accepted criteria of intelligence per se. There is at present an increasing tendency to add personality tests to such selection batteries in frank recognition of the fact that motivation and personal adjustment also in part determine an individual’s performance at the college level.

While some beginnings have been made in using intelligence tests for placement in industry, a particularly good example of their use can be seen in the military services. We shall use the U. S. Navy as our example. Shortly after arriving at his Naval training station, every recruit is given a group paper and pencil test known as the General Classification Test. Any recruit scoring below a certain level is then referred to the Psychiatric department for individual testing by a clinical psychologist to determine whether the low score is genuinely attributable to low intelligence or is the secondary result of some educational handicap (i.e. illiteracy, deficient schooling, etc.) or of some basic
psychopathosis. The Wechsler-Bellevue Scale is commonly used for this individual testing. If mental deficiency is established the recruit is separated from the service. If some educational handicap is at the basis of his poor performance he may be referred to a special remedial program. If emotional maladjustment or mental disorder is suspected he is referred to the psychiatrist. The group test results are also used for the later placement of the recruit. His score may be high enough to certify him for Naval service but not high enough to warrant further special educational training within the Navy. Should the test indicate an intelligence equal to the demands of a Naval Trade School he becomes eligible for such specialized training and may be sent to an electrician's school, carpenter's school, etc., depending upon the direction of his interests and special abilities. A still higher score is usually demanded for officers' training school.

The Clinical Use of Tests

Our use of the word clinical covers the broad field of psychological counselling in problems of personal adjustment as well as the narrower, psychiatric field of the actual mental disorders. Intelligence, as one aspect of an individual's abilities, characteristics, or traits, enters prominently into his ultimate adjustment to his environment. If life makes demands upon him that lie beyond the possibilities of his intelligence, emotional difficulties may ensue. Again, maladjustment may result if the demands of the environment are not up to the potentialities of the individual, if he does not get a chance to make use of his ability. Intelligence, while not a primary cause, is a contributing cause in many cases of delinquency. It may even dictate the general direction the symptoms may take in the mental disorders, as well as the amount of cooperation that the psychiatrist may expect during therapy, and the amount of insight that can be given the patient into the basic roots of his trouble. For all these reasons, some measure of an individual's intelligence is necessary in understanding his adjustment to life, and in directing the necessary therapeutic steps when this adjustment goes awry.

In our discussion of scatter we have already mentioned the diagnostic aid that tests furnish in understanding mental dis-
order. In showing what aspects of intelligence are differentially affected and how greatly they are disturbed, intelligence tests enable the clinical psychologist to furnish the psychiatrist with valuable clues toward grasping the type of disorder present.

The clinical aspects of intelligence testing cover its use in courts of law. The results of intelligence tests may contribute to the final decision in cases where responsibility, commitment to an institution, or the competency of a witness is in question. The test results are not admitted directly in evidence, but may furnish a basis for the opinion of an expert witness, i.e., a doctor or psychiatrist; or the court may call for testing to guide it in forming its own opinion. In some states, such as Massachusetts, routine testing of all individuals convicted of major crimes is provided for in order that the court may have this data available before passing sentence. The legal status of intelligence tests is still in question, and their legal use is still in its infancy, but the future should see a further development of such testing.

THE ABUSES OF TESTS

Intelligence tests, like any measuring instrument, may be abused. If the results of testing are to be reliable, a good test must be used, one that measures up to the criteria of objectivity, standardization, reliability, and validity discussed at the beginning of this article. Even when a good test is selected for use there are certain errors inherent in the statistical procedures used in constructing the test. Moreover a good test may be misused by a poor tester; and even when a good test is given by a good tester, the results may be misinterpreted. Finally a test may be abused by the person taking it, as happens in malingering. We will discuss the possible abuses of testing at greater length.

THE CHARACTERISTICS OF A GOOD TEST

First and foremost an intelligence test should be valid. It should agree with other known measures of intelligence, and this agreement should be experimentally ascertained and demonstrated as an integral part of the test's construction. Secondly, a test should be reliable, it should give the same results with repeated usage. In general it is hard to see how a test can be truly
valid and not reliable, but most psychologists prefer to treat these two important attributes separately. An intelligence test should be **objective**. It should provide for relatively simple, unequivocal answers, the correctness or incorrectness of which can be ascertained by inspection without subjective interpretation by the person doing the scoring. Finally, a good test should be **standardized**. It should provide for a fixed routine of administration which can be followed without variation. Improvisation in administration means that the conditions under which a test is given will vary from time to time and that the results in each case will no longer be strictly comparable. We thus lose the chance of comparing the individual’s performance with that of others in previous testing situations since such comparison is valid only if the test is always given in the same way under the same conditions.

Whether or not any given test fulfills the above requirements can always be ascertained by referring to the original sources describing the construction of the test. Most authors publish an abbreviated statement of such matters in the test manual which accompanies their test. The uninitiated layman, however, often has to accept the fulfillment of these requirements on faith. Fortunately, testing has developed to the point where there is universal professional acceptance of the criteria for a good test, and one can be reasonably certain that any widely used intelligence test published under reliable auspices will meet these criteria.

**Errors in Administration**

As we have just seen, if the results of testing are to be comparable from one situation to another, the test must always be administered in a standard fashion. This is provided for by an elaborate set of careful instructions provided with each test, and by the careful preliminary training of those persons qualified to administer tests. If an untrained person is allowed to give the tests, or if the standard procedures are not adhered to, the results cannot be considered adequate. Moreover, when the test is given, it must be given under favorable conditions. The subject should be comfortable and there should be no distractions present. One would not expect an individual standing on one leg, writing on the back of a book in the middle of a boiler...
factory to do justice to his true potentialities for intelligent performance. Finally, care and application are necessary for the correct scoring of tests. Even when an objective scoring stencil is used human errors may creep in. The increasing use of machine scoring, however, provides an accurate means of eliminating such errors.

Errors in Interpretation

In interpreting the results of an intelligence test we must remember that our measures are rough and that they all have a margin of error. While we use these tests in passing judgment upon the intelligence of a single individual, their validation was accomplished upon a group and involves statistical abstractions based upon group performance. As an illustration let us assume that we are constructing a new intelligence test. We might administer it to a group of children with a known mental age of eight years. If the test were valid we would expect the performance of these children to cluster around a certain score. They would all tend to answer a certain number of items correctly. We would then conclude that this score was typical for children with a mental age of eight years, and, in any table of norms we might establish, this score would be given as our standard for this age. We must remember, however, that our normative score for eight would express the average performance or central tendency of the group and that the actual scores obtained by the members of the group would be dispersed on either side of this central tendency. If the average score were 40, we might expect the actual scores to range from 35 to 45. The use of the average or mean score thus conceals the dispersion of the actual scores obtained and lends a false air of precision to our measure. In making predictions of the future performance on this test of a new group of children of the same mental age, we must remember that while we could predict an average score of 40 for the group, we could only predict that any one child would score somewhere between 35 and 45. The amount of dispersion around the group mean would thus represent the error inherent in our test. Such a margin of error is not large in most of our tests but its existence must be kept in mind.

Another source of error in interpretation enters when the
test is administered to *atypical* individuals. It is easily recognized that a paper and pencil test of intelligence which demands that the subject be able to read would not be suitable for measuring the intelligence of an illiterate. It is less clearly recognized that a pictorial, performance test may easily include materials which lie outside the experience of some culturally handicapped individual. We would not be fair in expecting a boy from some isolated, back-woods, mountain community to recognize the absence of threads from the base of an electric light bulb if he had never seen one before. The absence of a knob from a door is significant only if you have been brought up in an environment in which every door customarily has a knob.

Our standard intelligence tests have largely been constructed for use with individuals who have had the common educational opportunities of our democracy. There are many underprivileged groups, however, the members of which suffer from an educational deficiency which handicaps them on the usual intelligence test. This fact was obvious when large numbers of draftees from every strata of life were tested in the armed services in both World Wars. Illiteracy, educational handicap, and cultural handicap all may render invalid the results of an intelligence test. Before an individual's score can be accepted as an adequate index of his intelligence we must be sure that his educational and cultural background is representative of the group on which the test was originally validated and from whose performance the norms were derived.

The individual being tested must not only be educationally and culturally adequate to the demands of the testing situation but he must also be physically adequate. Extreme fatigue may lower his performance markedly. Illness can affect it, as can drug conditions. Overexcitement or emotional upset may also interfere. Some individuals can overcome these handicaps better than others, but it is seldom safe to assume that the results of testing are reliable unless the subject is healthy, calm, and in possession of his normal energy reserve.

Such conditions naturally are not fulfilled in the various mental disorders. The performance of a psychotic patient should be interpreted as a measure of performance during his disorder, and not as indicative of his previous level of intelligence or prognostic of his future level if he recovers. In addi-
tion to the specific deficits in performance that may trace directly to his mental disturbance (the basis for the diagnostic use of scatter mentioned above) such patients are often distractible, unresponsive, out of contact, and non-cooperative. Anyone who has ever labored to get test responses from a severely depressed patient will recognize these difficulties. Indeed there are those psychologists who claim that the poor test response of the psychotic is not attributable to any basic derangement of intellect, but to secondary factors of attention, concentration, rapport, etc. Deficit may also be expected in various types of organic brain disorders, such as arteriosclerosis, traumatic brain injury, cerebral paresis, etc. Experience has shown that in most cases of such deficit associated with mental disorder vocabulary is the least susceptible to deterioration of all our measures of intelligence. Where an approximation of previous intelligence level in such cases is desired, it is therefore wisest to depend upon some form of vocabulary tests.

Experience has also shown us that there is some tendency for the level of performance in many of the so-called psychopathic personalities to fall below the true potentialities of the individual. Such psychopathic personalities constitute the large body of asocial individuals whose peculiarities are perhaps best described as behavior disorders rather than true mental disorders accompanied by evidence of deterioration in thinking. No psychotic symptomatology is present in these individuals, and their outstanding characteristic is their asocial tendency expressed in their failure to conform to the demands of society. Oftentimes this asocial tendency will exhibit itself in a lack of cooperation with others, a lack of cooperation which is as evident in the testing situation as it is in the situations of ordinary life. Cooperation in the testing situation is usually spoken of as rapport. Rapport simply means that the individual being tested understands the situation, is sympathetic with the goal of the tester, and attempts to do his best on the test. Its presence is of course necessary with normal as well as abnormal individuals, in fact no test can be given successfully without it. Unless the individual is cooperative and motivated to do his best the test results cannot be considered as reliable.

See p. 24.
One further source of error in the interpretation of test results comes about in the failure to realize that intelligence is not the only factor in successful performance in life situations. The ability to succeed in solving the practical problems of life goes well beyond the mere possession of intelligence, and a high intelligence quotient by itself is no guarantee of success. This is equally true in education, in business, and in social matters. We know, for instance, that while there is a high relationship between intelligence and successful scholastic performance in grammar school this correlation drops increasingly as we progress through the higher educational institutions of high school, college, and professional school. It would seem that the more mature an individual becomes the more his ultimate performance depends upon factors of motivation and effort rather than upon intelligence alone. Intelligence is a contributing factor to success but not the only one.

MALINGERING

There remains to be discussed what we have called the abuse of a test by the person taking it. We refer to malingering or the conscious and willful attempt of the individual being tested to appear other than he really is. Under the proper conditions of test administration there is no opportunity to obtain a higher score by cheating. It is still possible for a person to commit deliberate errors in order to appear less intelligent. Simulation of mental deficiency was sometimes attempted in the armed services during the war, and may occasionally be met in civilian practice. Successful malingering is exceedingly difficult, however, and it is not as easy to “beat the game” as some people believe. The mind of the mental deficient follows certain definite patterns in its functioning and these patterns express themselves in typical characteristic responses on an intelligence test. The feebleminded individual tends to fail certain items and to fail them in certain ways. When we observe his test behavior we find it follows a definite predictable pattern which is different from that of the malingerer. In general the malingerer will not fail the same items that the mental deficient does and when he does fail the same items he fails in a different fashion.
The difference is seen in the responses to the question "If 8 boys club together and pay 2 dollars for the use of a room, how much should each pay?" The two most common answers that mental deficient gives are $4 and $16. He apparently isolates the 8 and 2 as the necessary elements in the problem, and realizes that something more complex than addition or subtraction is called for. Unable to divide 2 by 8, he either divides the 8 by 2 and gets 4, or multiplies them and gets 16. The malingerer, on the other hand, successfully accomplishes the basic problem of dividing 2 by 8 and then deliberately distorts the result by a few cents offering a final answer of 23¢, 26¢, etc. This type of error is readily recognized and leads to his detection.

Once the characteristic test behavior of the true mental deficient and the malingerer has been observed and plotted, the quantitative and qualitative differences that appear render the detection of malingering relatively easy. This is borne out by the experience of both the Army and the Navy.

**Conclusion**

Here again, in the detection of malingering, we have another instance of the value of applying the objective procedures of scientific observation and classification. In conclusion, let us reiterate that intelligence testing is no esoteric practice based upon the *a priori* judgment of the psychologist. Intelligence tests result from the orderly application of scientific procedures to the problem of differential classification. The careful observation of human behavior reveals that in a selected series of situations, as represented by our test items, different levels of performance will be found which are characteristic of different levels of intelligence. When this correspondence stands up under repeated investigation the respective levels of performance are then taken as indicative of the corresponding levels of intelligence. In the future any individual performing at a certain level on our tests is assumed to possess the corresponding level of intelligence. We are dealing with the development of scientific prediction in a field where such prediction is of immense value.

---


to society. Like any other instrument of measurement intelligence tests have their limitations, but a good test in skilled hands will give a reliable measure. The practical answer to the value of intelligence testing lies in its current acceptance and widespread use.
EDITORIAL BOARD
1946-1947

FACULTY OF THE COLLEGE OF LAW, EX OFFICIO
Roy Moreland, Faculty Editor

Selby Hurst, Student Editor
Durward Weldon, Managing Editor

JAMES ADKINS
JAMES C. BROCK
WILLIAM COLDIRON
WILLIAM GILBREATH

DOYLE INMAN
JOHN MORROW KINNAIRD
ARNETT MANN
POLLARD WHITE

ADVISORY BOARD FROM THE STATE BAR ASSOCIATION

Rosanna A. Blake, Washington, D. C., Chairman
Term Expires, 1947

James W. Cammack, Frankfort
William H. Conley, Carlisle
Edward A. Dodd, Louisville
Elson S. Dummit, Frankfort
Richard H. Hill, Louisville
Charles S. Landrum, Lexington
Frank L. McCarthy, Lexington
Leslie W. Morris, Frankfort
J. Owen Reynolds, Lexington
 Mildred Robards, Lexington
Colvin Rouse, Versailles
Ira Stephenson, Williamstown

Gregory Bruce, Washington, D. C.
T. C. Carroll, Shepherdsville
J. Donald Dinning, Louisville
Nathan Elliott, Jr., Lexington
A. E. Funk, Frankfort
Roy Helm, Hazard
Robert F. Houlihan, Lexington
Ben Kessinger, Lexington
Lee B. Lanter, Williamstown
C. E. Rankin, Harrodsburg
J. J. Tye, Barbourville
C. C. Wells, Hazard